JC17 Rec'd PCT/PTO 0 8 MAY 2001

U.S. DEPARTMENT OF COMMERCE PATENT AND TRADEMARK OFFICE **TRANSMITTAL LETTER TO THE UNITED STATES DESIGNATED/ELECTED OFFICE (DO/EO/US) CONCERNING A FILING UNDER 35 USC 371**		ATTORNEY DOCKET NO. 211163 U.S. APPLICATION NO. 09/831262
INTERNATIONAL APPLICATION NO. PCT/GB99/03737	INTERNATIONAL FILING DATE 9 November 1999	PRIORITY DATE CLAIMED 9 November 1998
TITLE OF INVENTION DATA CLASSIFICATION APPARATUS AND METHOD THEREOF		
APPLICANT(S) FOR DO/EO/US Gammerman et al.		

Applicant herewith submits to the United States Designated/Elected Office (DO/EO/US) the following items and other information:

1. ☒ This is a **FIRST** submission of items concerning a filing under 35 USC 371.
2. ☐ This is a **SECOND** or **SUBSEQUENT** submission of items concerning a filing under 35 USC 371.
3. ☒ This is an express request to begin national examination procedures (35 USC 371(f)).
4. ☒ The US has been elected by the expiration of 19 months from the priority date (PCT Article 31).
5. ☒ A copy of the International Application as filed (35 USC 371(c)(2))
   a. ☐ is attached hereto (required only if not communicated by the International Bureau).
   b. ☒ has been communicated by the International Bureau.
   c. ☐ is not required, as the application was filed in the United States Receiving Office (RO/US).
6. ☐ An English language translation of the International Application as filed (35 USC 371(c)(2)).
7. ☒ Amendments to the claims of the International Application under PCT Article 19 (35 USC 371(c)(3))
   a. ☐ are attached hereto (required only if not communicated by the International Bureau).
   b. ☐ have been communicated by the International Bureau.
   c. ☐ have not been made; however, the time limit for making such amendments has NOT expired.
   d. ☒ have not been made and will not be made.
8. ☐ An English language translation of the amendments to the claims under PCT Article 19 (35 USC 371(c)(3)).
9. ☐ An oath or declaration of the inventor(s) (35 USC 371(c)(4)).
10. ☐ An English language translation of the annexes to the International Preliminary Examination Report under PCT Article 36 (35 USC 371(c)(5)).
11. Nucleotide and/or Amino Acid Sequence Submission
    a. ☐ Computer Readable Form (CRF)
    b. Specification Sequence Listing on:
       i. ☐ CD-ROM or CD-R (2 copies); or
       ii. ☐ Paper Copy
    c. ☐ Statement verifying identity of above copies

Items 12 to 19 below concern other document(s) or information included:

12. ☐ An Information Disclosure Statement under 37 CFR 1.97 and 1.98.
    ☐ Form PTO-1449
    ☐ Copies of Listed Documents
13. ☐ An assignment for recording. A separate cover sheet in compliance with 37 CFR 3.28 and 3.31 is included.
14. ☐ A FIRST preliminary amendment.
    ☐ A SECOND or SUBSEQUENT preliminary amendment.
15. ☐ A substitute specification.
16. ☐ A change of power of attorney and/or address letter.
17. ☒ Application Data Sheet Under 37 CFR 1.76
18. ☒ Return Receipt Postcard
19. ☐ Other items or information:

U.S. APPLICATION NO. 09/831262	INTERNATIONAL APPLICATION NO. PCT/GB99/03737	ATTORNEY DOCKET NO. 211163

20. ☒ The following fees are submitted:	CALCULATIONS	PTO USE ONLY
Basic National Fee (37 CFR 1.492(a)(1)-(5)): Neither international preliminary examination fee (37 CFR 1.482) nor international search fee (37 CFR 1.445(a)(2)) paid to USPTO and International Search Report not prepared by the EPO or JPO .... $1,000.00 International preliminary examination fee (37 CFR 1.482) not paid to USPTO but International Search Report prepared by the EPO or JPO .... $ 860.00 International preliminary examination fee (37 CFR 1.482) not paid to USPTO, but international search fee (37 CFR 1.445(a)(2))paid to USPTO .... $ 710.00 International preliminary examination fee paid to USPTO (37 CFR 1.482) but all claims did not satisfy provisions of PCT Article 33(1)-(4)........................ $ 690.00 International preliminary examination fee paid to USPTO (37 CFR 1.482) and all claims satisfied provisions of PCT Article 33(1) to (4) .... $ 100.00		
ENTER APPROPRIATE BASIC FEE AMOUNT=	$860.00	
Surcharge of $130.00 for furnishing the National fee or oath or declaration later than ☐ 20 ☐ 30 months from the earliest claimed priority date	$	

CLAIMS	NUMBER FILED	NUMBER EXTRA	RATE		
Total Claims	9 -20=	0	x $ 18.00	$	
Independent Claims	3 - 3 =	0	x $ 80.00	$	
☐ Multiple Dependent Claim(s) (if applicable)			+$270.00	$	

TOTAL OF ABOVE CALCULATIONS=	$860.00	
☐ Applicant claims small entity status. See 37 CFR 1.27. The fees indicated above are reduced by 1/2.	$	
SUBTOTAL=	$860.00	
Processing fee of **$130.00** for furnishing English Translation later than ☐ 20 ☐ 30 months from the earliest claimed priority date.	$	
TOTAL NATIONAL FEE=	$860.00	
Fee for recording the enclosed assignment. The assignment must be accompanied by an appropriate cover sheet. **$40.00** per property +	$	
TOTAL FEE ENCLOSED=	$860.00	
	Amount to be: refunded	$
	charged:	$

a. ☒ A check in the amount of $860.00 to cover the above fee is enclosed.

b. ☐ Please charge Deposit Account No. 12-1216 in the amount of $ to cover the above fees. A duplicate copy of this sheet is enclosed.

c. ☒ The Commissioner is hereby authorized to charge any additional fees which may be required, or credit any overpayment to Deposit Account No. 12-1216. A duplicate copy of this sheet is enclosed.

NOTE: Where an appropriate time limit under 37 CFR 1.494 or 1.495 has not been met, a petition to revive (37 CFR 1.137(a) or (b)) must be filed and granted to restore the application to pending status.

SEND ALL CORRESPONDENCE TO:

Customer Number: 23460

23460

PATENT TRADEMARK OFFICE

Dennis R. Schlemmer, Registration No. 24,703
One of the Attorneys for Applicant(s)

Date: May 8, 2001

JC18 Rec'd PCT/PTO 0 8 MAY 2001

U.S. APPLICATION NO. 09/831262	INTERNATIONAL APPLICATION NO. PCT/GB99/03737	ATTORNEY DOCKET NO. 211163

CERTIFICATION UNDER 37 CFR 1.10

"Express Mail" Label Number: EL304645035US

Date of Deposit: May 8, 2001

I hereby certify that this express request to begin national examination procedures under 35 USC 371(f) of the International Patent Application referenced above, including all of the items listed thereon as enclosures, is being deposited with the United States Postal Service "Express Mail Post Office to Addressee" Service under 37 CFR 1.10 on the date indicated above and is addressed to Box PCT, Assistant Commissioner for Patents, Attention: DO/EO/US, Washington, D.C. 20231.

WILMA DEL NAGRO

Printed Name of Person Signing:

Wilma Del Nagro

Signature

09/831262

JC18 Rec'd PCT/PTO 08 MAY 2001

PATENT

Attorney Docket No. 211163

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of:

Alex Gammerman
Volodya Vovk

Application No.

Filed:

Art Unit: Unassigned

Examiner: Unassigned

For: DATA CLASSIFICATION APPARATUS AND METHOD THEREOF

PRELIMINARY AMENDMENT

Commissioner for Patents
Washington, D.C. 20231

Dear Sir:

Prior to the examination of the above-identified patent application, please enter the following amendments and consider the following remarks.

AMENDMENTS

IN THE CLAIMS:

Please cancel claims 1-9 without prejudice, and add new claims 10-18 as follows:

10. Data classification apparatus comprising:

an input device for receiving a plurality of training classified examples and at least one unclassified example;

a memory for storing said classified and unclassified examples;

an output terminal for outputting a predicted classification for said at least one unclassified example; and

a processor for identifying the predicted classification of said at least one unclassified example

wherein the processor includes:

classification allocation means for allocating potential classifications to each said unclassified example and for generating a plurality of classification sets, each said classification set containing said plurality of training classified examples and said at least one unclassified example with its said allocated potential classification;

assay means for determining a strangeness value valid under the iid assumption for each said classification set;

a comparative device for selecting the classification set to which the most likely allocated potential classification for said at least one unclassified example belongs, wherein said predicted classification output by the output terminal is said most likely allocated classification according to said strangeness values assigned by said assay means; and

a strength of prediction monitoring device for determining a confidence value for said predicted classification on the basis of said strangeness value assigned by said assay means to one of said classification sets to which the second most likely allocated potential classification of said at least one unclassified example belongs.

11. Data classification apparatus as claimed in claim 10, wherein said processor further includes an example valuation device which determines individual strangeness values for each said training classified example and said at least one unclassified example having an allocated potential classification.

12. Data classification apparatus as claimed in claim 11, wherein Lagrange multipliers are used to determine said individual strangeness values.

13. Data classification apparatus as claimed in claim 11, wherein said assay means determines a strangeness value for each said classification set in dependence on said individual strangeness values of each said example.

14. Data classification apparatus comprising:

an input device for receiving a plurality of training classified examples and at least one unclassified example;

a memory for storing said classified and unclassified examples;

stored programs including an example classification program;

an output terminal for outputting a predicted classification for said at least one unclassified example; and

a processor controlled by said stored programs for identifying the predicted classification of said at least one unclassified example,

wherein said processor includes:

classification allocation means for allocating potential classifications to each said unclassified example and for generating a plurality of classification sets, each said classification set containing said plurality of training classified examples and said at least one unclassified example with its allocated potential classification;

assay means for determining a strangeness value valid under the iid assumption for each said classification set;

a comparative device for selecting the classification set to which the most likely allocated potential classification for said at least one unclassified example belongs, wherein the predicted classification output by said output terminal is the most likely allocated potential classification according to said strangeness values assigned by said assay means and

a strength of prediction monitoring device for determining a confidence value for said predicted classification on the basis of said strangeness value assigned by said assay means to one of said classification sets to which the second most likely allocated potential classification of said at least one unclassified example belongs.

15. A data classification method comprising:

inputting a plurality of training classified examples and at least one unclassified example;

identifying a predicted classification of said at least one unclassified example which includes,

allocating potential classifications to each said unclassified example;

generating a plurality of classification sets, each said classification set containing said plurality of training classified examples and said at least one unclassified example with its allocated potential classification;

determining a strangeness value valid under the iid assumption for each said classification set;

selecting the said classification set to which the most likely allocated potential classification for said at least one unclassified example belongs, wherein said predicted classification is the most likely allocated potential classification in dependence on said strangeness values;

determining a confidence value for said predicted classification on the basis of the strangeness value assigned to one of said classification sets to which the second most likely allocated potential classification for said at least one unclassified example belongs; and

outputting said predicted classification for said at least one unclassified example and said confidence value for said predicted classification.

16. A data classification method as claimed in claim 15, further including determining individual strangeness values for each said training classified example and said at least one unclassified example having an allocated potential classification.

17. A data classification method as claimed in claim 15, wherein said selected classification set is selected without the application of any general rules determined from the said training set.

18. A data carrier on which is stored a classification program for classifying data by performing the following steps:

In re Application of: Gammerman et al.
Attorney Docket No. 211163

generating a plurality of classification sets, each said classification set containing a plurality of training classified examples and at least one unclassified example that has been allocated a potential classification;

determining a strangeness value valid under the iid assumption for each said classification set;

selecting the classification set to which the most likely allocated potential classification for the said at least one unclassified example belongs, wherein the predicted classification is the most likely allocated potential classification in dependence on said strangeness values; and

determining a confidence value for said predicted classification on the basis of said strangeness value assigned to one of said classification sets to which the second most likely allocated potential classification for said at least one unclassified example belongs.

Respectfully submitted,

Dennis R. Schlemmer, Reg. No. 24,703
One of the Attorneys for Applicant(s)
LEYDIG, VOIT & MAYER, LTD.
Two Prudential Plaza, Suite 4900
180 North Stetson
Chicago, Illinois 60601-6780
(312) 616-5600 (telephone)
(312) 616-5700 (facsimile)

Date: May 8, 2001

09/831262
Rec'd PCT/PTO 08 MAY 2001
PCT/GB99/03737

DATA CLASSIFICATION APPARATUS AND METHOD THEREOF

BACKGROUND OF THE INVENTION

The present invention relates to data classification apparatus and an automated method of data classification thereof that provides a universal measure of confidence in the predicted classification for any unknown input. Especially, but not exclusively, the present invention is suitable for pattern recognition, e.g. optical character recognition.

In order to automate data classification such as pattern recognition the apparatus, usually in the form of a computer, must be capable of learning from known examples and extrapolating to predict a classification for new unknown examples. Various techniques have been developed over the years to enable computers to perform this function including, inter alia, discriminant analysis, neural networks, genetic algorithms and support vector machines. These techniques usually originate in two fields: machine learning and statistics.

Learning machines developed in the theory of machine learning often perform very well in a wide range of applications without requiring any parametric statistical assumptions about the source of data (unlike traditional statistical techniques); the only assumption made is the iid assumption (the examples are generated from the same probability distribution independently of each other). A new approach to machine learning is described in US5640492, where mathematical optimisation techniques are used for classifying new examples. The advantage of the learning machine described in US5640492 is that it can be used for solving extremely high-dimensional problems which are infeasible for the previously known learning machines.

A typical drawback of such techniques is that the techniques do not provide any measure of confidence in the predicted classification output by the apparatus. A typical user of such data classification apparatus just hopes that the accuracy of the results from previous analyses using benchmark datasets is representative of the results to be obtained from the

analysis of future datasets.

Other options for the user who wants to associate a measure of confidence with new unclassified examples include performing experiments on a validation set, using one of the known cross-validation procedures, and applying one of the theoretical results about the future performance of different learning machines given their past performance. None of these confidence estimation procedures though provides any practicable means for assessing the confidence of the predicted classification for an individual new example. Known confidence estimation procedures that address the problem of assessing the confidence of a predicted classification for an individual new example are ad hoc and do not admit interpretation in rigorous terms of mathematical probability theory.

Confidence estimation is a well-studied area of both parametric and non-parametric statistics. In some parts of statistics the goal is classification of future examples rather than of parameters of the model, which is relevant to the need addressed by this invention. In statistics, however, only confidence estimation procedures suitable for low-dimensional problems have been developed. Hence, to date mathematically rigorous confidence assessment has not been employed in high-dimensional data classification.

SUMMARY OF THE INVENTION

The present invention provides a new data classification apparatus and method that can cope with high-dimensional classification problems and that provides a universal measure of confidence, valid under the iid assumption, for each individual classification prediction made by the new data classification apparatus and method.

The present invention provides data classification apparatus comprising: an input device for receiving a plurality of training classified examples and at least one unclassified example; a memory for storing the classified and unclassified examples; an output terminal for outputting a predicted classification for the at least one unclassified example; and a processor for identifying the predicted classification of the at least one

unclassified example wherein the processor includes: classification allocation means for allocating potential classifications to each unclassified example and for generating a plurality of classification sets, each classification set containing the plurality of training classified examples and the at least one unclassified example with its allocated potential classification; assay means for determining a strangeness value for each classification set; and a comparative device for selecting a classification set containing the most likely allocated potential classification for at least one unclassified example, whereby the predicted classification output by the output terminal is the most likely allocated potential classification, according to the strangeness values assigned by the assay means.

In the preferred embodiment the processor further includes a strength of prediction monitoring device for determining a confidence value for the predicted classification on the basis of the strangeness value of a set containing the at least one unclassified example with the second most likely allocated potential classification.

With the present invention the conventional data classification technique of induction learning and then deduction for new unknown data vectors is supplanted by a new transduction technique that avoids the need to identify any all encompassing general rule. Thus, with the present invention no multidimensional hyperplane or boundary is identified. The training data vectors are used directly to provide a predicted classification for unknown data vectors. In other words, the training data vectors implicitly drive classification prediction for an unknown data vector.

It is important to note that with the present invention the measure of confidence is valid under the general iid assumption and the present invention is able to provide measures of confidence for even very high dimensional problems.

Furthermore, with the present invention more than one unknown data vector can be classified and a measure of confidence generated simultaneously.

In a further aspect the present invention provides data classification

apparatus comprising: an input device for receiving a plurality of training classified examples and at least one unclassified example; a memory for storing the classified and unclassified examples; stored programs including an example classification program; an output terminal for outputting a predicted classification for the at least one unclassified example; and a processor controlled by the stored programs for identifying the predicted classification of the at least one unclassified example wherein the processor includes: classification allocation means for allocating potential classifications to each unclassified example and for generating a plurality of classification sets, each classification set containing the plurality of training classified examples and the at least one unclassified example with its allocated potential classification; assay means for determining a strangeness value for each classification set; and a comparative device for selecting a classification set containing the most likely allocated potential classification for the at least one unclassified example, whereby the predicted classification output by the output terminal is the most likely allocated potential classification, according to the strangeness values assigned by the assay means.

In a third aspect the present invention provides a data classification method comprising:

inputting a plurality of training classified examples and at least one unclassified example;

identifying a predicted classification of the at least one unclassified example which includes

allocating potential classifications to each unclassified example;

generating a plurality of classification sets each containing the plurality of training classified examples and the at least one unclassified example with an allocated potential classification;

determining a strangeness value for each classification set; and

selecting, according to the assigned strangeness values, a

classification set containing the most likely allocated potential classification; and outputting the predicted classification for the at least one unclassified example whereby the predicted classification output by an output terminal is the most likely allocated potential classification.

It will, of course, be appreciated that the above method and apparatus may be implemented in a data carrier on which is stored a classification program.

BRIEF DESCRIPTION OF THE DRAWINGS

An embodiment of the present invention will now be described by way of example only with reference to the accompanying drawings, in which:

Figure 1 is a schematic diagram of data classification apparatus in accordance with the present invention;

Figure 2 is a schematic diagram of the operation of data classification apparatus of Figure 1;

Figure 3 is a table showing a set of training examples and unclassified examples for use with a data classifier in accordance with the present invention; and

Figure 4 is a tabulation of experimental results where a data classifier in accordance with the present invention was used in character recognition.

DESCRIPTION OF PREFERRED EMBODIMENT

In Figure 1 a data classifier 10 is shown generally consisting of an input device 11, a processor 12, a memory 13, a ROM 14 containing a suite of programs accessible by the processor 12 and an output terminal 15. The input device 11 preferably includes a user interface 16 such as a keyboard or other conventional means for communicating with and inputting data to the processor 12 and the output terminal 15 may be in the form of a display monitor or other conventional means for displaying information to a user. The output terminal 15 preferably includes one or more output ports for connection to a printer or other network device. The data classifier 10 may be embodied in an Application Specific Integrated

Circuit (ASIC) with additional RAM chips. Ideally, the ASIC would contain a fast RISC CPU with an appropriate Floating Point Unit.

To assist in an understanding of the operation of the data classifier 10 in providing a prediction of a classification for unclassified (unknown) examples, the following is an explanation of the mathematical theory underlying its operation.

Two sets of examples (data vectors) are given: the training set consists of examples with their classifications (or *classes)* known and a test set consisting of unclassified examples. In Figure 3, a training set of five examples and two test examples are shown, where the unclassified examples are images of digits and the classification is either 1 or 7.

The notation for the size of the training set is l and, for simplicity, it is assumed that the test set of examples contains only one unclassified example. Let (X,A) be the measurable space of all possible unclassified examples (in the case of Figure 3, X might be the set of all 16 *x* 16 grey-scale images) and (Y,B) be the measurable space of classes (in the case of Figure 3, Y might be the 2-element set {1, 7}). Y is typically finite.

The confidence prediction procedure is a family $\{f_\beta:\beta\in(0, 1]\}$ of measurable mappings $f_\beta:(X \times Y)^l \times X\rightarrow B$ such that:

1. For any confidence level β (in data classification typically we are interested in β close to 1) and any probability distribution P in $X \times Y$, the probability that

$$y_{l+1} \in f_\beta(x_1, y_1, \ldots, x_l, y_l, x_{l+1})$$

is at least β, where $(x_1, y_1), \ldots, (x_l, y_l), (x_{l+1}, y_{l+1})$ are generated independently from P.

2. If $\beta_1 < \beta_2$, then, for all $(x_1, y_1, \ldots, x_l, y_l, x_{l+1}) \in (X \times Y)^l \times X$,

$$f_{\beta_1}(x_1, y_1, \ldots, x_l, y_l, x_{l+1}) \subseteq f_{\beta_2}(x_1, y_1, \ldots, x_l, y_l, x_{l+1})$$

The assertion implicit in the prediction $f_{\beta_1}(x_1, y_1, \ldots, x_l, y_l, x_{l+1})$ is that the true label y_{l+1} will belong to $f_{\beta_1}(x_1, y_1, \ldots, x_l, y_l, x_{l+1})$. Item 1 requires that the prediction given by f_β should be correct with probability at least β, and item 2 requires that the family $\{f_\beta\}$ should be consistent: if some label y for the (l+1)th example is allowed at confidence level β_1, it should also be allowed at any confidence level $\beta_2 > \beta_1$.

A typical mode of use of this definition is that some conventional value of β such as 95% or 99%, is chosen in advance, after which the function f_β is used for prediction. Ideally, the prediction region output by f_β will contain only one classification.

An important feature of the data classification apparatus is defining f_β in terms of solutions α_i, i=1, ..., l+1, to auxiliary optimisation problems of the kind outlined in US5640492, the contents of which is incorporated herein by reference. Specifically, we consider $|Y|$ completions of our data

$$(x_1, y_1), \ldots, (x_l, y_l), x_{l+1}$$

the completion y, $y \in Y$, is

$$(x_1, y_1), \ldots, (x_l, y_l), (x_{l+1}, y)$$

(so in all completions every example is classified).

With every completion

$$(x_1, y_1), \ldots, (x_l, y_l), (x_{l+1}, y_{l+1})$$

(for notational convenience we write y_{l+1} in place of y here) is associated the optimisation problem

$$\frac{1}{2}(w \cdot w) + C\left(\sum_{i=1}^{l+1} \xi_i\right) \longrightarrow \min \qquad (1)$$

(where C is a fixed positive constant)

subject to the constraints

$$y_i((x_i \cdot w) + b) \geq \xi_i, i = 1, \ldots, l+1 \qquad (2)$$

This problem involves non-negative variables $\xi_i \geq 0$, which are called *slack*

variables. If the constant C is chosen too large, the accuracy of solution can become unacceptably poor; C should be chosen as large as possible in the range in which the numerical accuracy of solution remains reasonable. (When the data is linearly separable, it is even possible to set C to infinity, but since it is rarely if ever possible to tell in advance that all completions will be linearly separable, C should be taken large but finite.)

The optimisation problem is transformed, via the introduction of Lagrange multipliers α_i, i=1, ..., l+1, to the dual problem: find α_i from

$$\sum_{i=1}^{l+1} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l+1} y_i y_j \alpha_i \alpha_j (x_i \cdot x_j) \longrightarrow \max \qquad (3)$$

under the "box" constraints

$$0 \le \alpha_i \le C,\ i=1, 2, ..., l+1 \qquad (4)$$

The unclassified examples are represented, it is assumed, as the values taken by n numerical attributes and so $X = \boldsymbol{R}^n$.

This quadratic optimisation problem is applied not to the attribute vectors x_i themselves, but to their images $\boldsymbol{V}(x_i)$ under some predetermined function $\boldsymbol{V}: X \rightarrow \boldsymbol{H}$ taking values in a Hilbert space, which leads to replacing the dot product $x_i \cdot x_j$ in the optimisation problem (3)—(4) by the kernel function

$$\boldsymbol{K}(x_i, x_j) = \boldsymbol{V}(x_i) \cdot \boldsymbol{V}(x_j)$$

The final optimisation problem is, therefore,

$$\sum_{i=1}^{l+1} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l+1} y_i y_j \alpha_i \alpha_j \boldsymbol{K}(x_i, x_j) \rightarrow \max$$

under the "box" constraints

$$0 \le \alpha_i \le C,\ i=1, 2, ..., l+1$$

this quadratic optimisation problem can be solved using standard packages.

The Lagrange multiplier α_i, $i \in \{1, ..., l+1\}$, reflects the "strangeness" of the example (x_i, y_i); we expect that α_{l+1} will be large in the wrong completions.

For $y \in Y$, define

$$d(y) := \frac{|\{i : \alpha_i \geq \alpha_{l+1}\}|}{l+1}$$

therefore $d(y)$ is the p-value associated with the completion y (y being an alternative notation for y_{l+1}). The confidence prediction function f, which is at the core of this invention, can be expressed as

$$f_\beta(x_1, y_1, \ldots, x_l, y_l, x_{l+1}) := \{y : d(y) > 1 - \beta\}$$

The most interesting case is where the prediction set given by f_β is a singleton; therefore, the most important features of the confidence prediction procedure $\{f_\beta\}$ at the data $(x_1, y_1), \ldots, (x_l, y_l), x_{l+1}$ are:

- the largest $\beta=\beta_0$ for which $f_\beta((x_1, y_1), \ldots, (x_l, y_l), x_{l+1})$ is a singleton (assuming such a β exists);
- the classification $\mathbf{F}((x_1, y_1), \ldots, (x_l, y_l), x_{l+1})$ defined to be that $y \in Y$ for which $f_{\beta_0}((x_1, y_1, \ldots, (x_l, y_l), x_{l+1})$ is $\{y\}$.

$\mathbf{F}((x_1, y_1), \ldots, (x_l, y_l), x_{l+1})$ defined in this way is called the f-optimal prediction algorithm; the corresponding β_0 is called the confidence level associated with **F**.

Another important feature of the confidence estimation function $\{f_\beta\}$ at the data $(x_1, y_1), \ldots, (x_l, y_l), x_{l+1}$ is the largest $\beta=\beta_*$ for which $f_\beta((x_1, y_1), \ldots, (x_l, y_l), x_{l+1})$ is the empty set. We call 1-β_* the credibility of the data set $(x_1, y_1), \ldots, (x_l, y_l), x_{l+1}$; it is the p-value of a test for checking the iid assumption. Where the credibility is very small, either the training set $(x_1, y_1), \ldots, (x_l, y_l)$ or the new unclassified example x_{l+1} are untypical, which renders the prediction unreliable unless the confidence level is much closer to 1, than is 1-β_*. In general, the sum of the confidence and credibility is between 1 and 2; the success of the prediction is measured by how close this sum is to 2.

With the data classifier of the present invention operated as

described above, the following menus or choices may be offered to a user:

1. Prediction and Confidence
2. Credibility
3. Details.

A typical response to the user's selection of choice 1 might be prediction: 4, confidence: 99%, which means that 4 will be the prediction output by the f-optimal **F** and 99% is the confidence level of this prediction. A typical response to choice 2 might be credibility: 100%, which gives the computed value of credibility. A typical response to choice 3 might be:

0	1	2	3	4	5	6	7	8	9
0.1%	1%	0.2%	0.4%	100%	1.1%	0.6%	0.2%	1%	1%

the complete set of p-values for all possible completions. The latter choice contains the information about **F**$((x_1, y_1),\ldots,(x_l, y_l), x_{l+1})$ (the character corresponding to the largest p-value), the confidence level (one minus the second largest p-value) and the credibility (the largest p-value).

This mode of using the confidence prediction function f is not the only possible mode: in principle it can be combined with any prediction algorithm. If **G** is a prediction algorithm, with its prediction y:=**G**$((x_1, y_1),\ldots,(x_l, y_l), x_{l+1})$ we can associate the following measure of confidence:

$$c(y) := \max\{\beta : f_\beta(x_1, y_1, \ldots, x_l, y_l, x_{l+1}) \subseteq \{y\}\}$$

The prediction algorithm **F** described above is the one that optimises this measure of confidence.

The table shown in Figure 4 contains the results of an experiment in character recognition using the data classifier of the present invention. The table shows the results for a test set of size 10, using a training set of size 20 (not shown). The kernel used was $\boldsymbol{K}(x, y) = (x \cdot y)^3 / 256$.

It is contemplated that some modifications of the optimisation problem set out under equations (l) and (2) might have certain advantages, for example,

$$\frac{1}{2}(w \cdot w) + C\left(\sum_{i=1}^{l+1} \xi_i^2\right) \rightarrow \min,$$

subject to the constraints

$$y_i((x_i \cdot w) + b) = 1 - \xi_i, i = 1, \ldots, l+1$$

It is further contemplated that the data classifier described above may be particularly useful for predicting the classification of more than one example simultaneously; the test statistic used for computing the p-values corresponding to different completions might be the sum of the ranks of αs corresponding to the new examples (as in the Wilcoxon rank-sum test).

In practice, as shown in Figure 2, a training dataset is input 20 to the data classifier. The training dataset consists of a plurality of data vectors each of which has an associated known classification allocated from a set of classifications. For example, in numerical character recognition, the set of classifications might be the numerical series 0—9. The set of classifications may separately be input 21 to the data classifier or may be stored in the ROM 14. In addition, some constructive representation of the measurable space of the data vectors may be input 22 to the data classifier or again may be stored in the ROM 14. For example, in the case of numerical character recognition the measurable space might consist of 16x16 pixellated grey-scale images. Where the measurable space is already stored in the ROM 14 of the data classifier, the interface 16 may include input means (not shown) to enable a user to input adjustments for the stored measurable space. For example, greater definition of an image may be required in which case the pixellation of the measurable space could be increased.

One or more data vectors for which no classification is known are also input 23 into the data classifier. The training dataset and the unclassified data vectors along with any additional information input by the user are then fed from the input device 11 to the processor 12.

Firstly, each one of the one or more unclassified data vectors is provisionally individually allocated 24 a classification from the set of

classifications. An individual strangeness value α_i is then determined 25 for each of the data vectors in the training set and for each of the unclassified data vectors for which a provisional classification allocation has been made. A classification set is thus generated containing each of the data vectors in the training set and the one or more unclassified data vectors with their allocated provision classifications and the individual strangeness values α_i for each data vector. A plurality of such classification sets is then generated with the allocated provisional classifications of the unclassified data vectors being different for each classification set.

Computation of a single strangeness value, the p-value, for each classification set containing the complete set of training data vectors and unclassified vectors with their current allocated classification is then performed 26, on the basis of the individual strangeness values α_i determined in the previous step. This p-value and the associated set of classifications is transferred to the memory 13 for future comparison whilst each of the one or more unclassified data vectors is provisionally individually allocated with the same or a different classification. The steps of calculating individual strangeness values 25 and the determination of a p-value 26 are repeated in each iteration for the complete set of training data vectors and the unclassified data vectors, using different classification allocations for the unclassified data vectors each time. This results in a series of p-values being stored in the memory 13 each representing the strangeness of the complete set of data vectors with respect to unique classification allocations for the one or more unclassified data vectors.

The p-values stored in the memory are then compared 27 to identify the maximum p-value and the next largest p-value. Finally, the classification set of data vectors having the maximum p-value is supplied 28 to the output terminal 15. The data supplied to the output terminal may consist solely of the classification(s) allocated to the unclassified data vector(s), which now represents the predicted classification, from the classification set of data vectors having the maximum p-value.

Furthermore, a confidence value for the predicted classification is

generated 29. The confidence value is determined based on the subtraction of the next largest p-value from 1. Hence, if the next largest p-value is large, the confidence of the predicted classification is small and if the next largest p-value is small, the confidence value is large. Choice 1 referred to earlier, provides a user with predicted classifications for the one or more unknown data vectors and the confidence value.

Where an alternative prediction algorithm is to be used, the confidence value will be computed by subtracting from 1 the largest p-value for the sets of training data vectors and new vectors classified differently from the predicted (by the alternative method) classification.

Additional information in the form of the p-values for each of the sets of data vectors with respect to the individual allocated classifications may also be supplied (choice 3) or simply the p-value for the predicted classification (choice 2).

With the data classifier and method of data classification described above, a universal measure of the confidence in any predicted classification of one or more unknown data vectors is provided. Moreover, at no point is a general rule or multidimensional hyperplane extracted from the training set of data vectors. Instead, the data vectors are used directly to calculate the strangeness of a provisionally allocated classification(s) for one or more unknown data vectors.

While the data classification apparatus and method have been particularly shown and described with reference to the above preferred embodiment, it will be understood by those skilled in the art that various modifications in form and detail may be made therein without departing from the scope and spirit of the invention. Accordingly, modifications such as those suggested above, but not limited thereto, are to be considered within the scope of the invention.

CLAIMS

1. Data classification apparatus comprising:

an input device for receiving a plurality of training classified examples and at least one unclassified example;

a memory for storing the classified and unclassified examples;

an output terminal for outputting a predicted classification for the at least one unclassified example; and

a processor for identifying the predicted classification of the at least one unclassified example

wherein the processor includes:

classification allocation means for allocating potential classifications to each unclassified example and for generating a plurality of classification sets, each classification set containing the plurality of training classified examples and the at least one unclassified example with its allocated potential classification;

assay means for determining a strangeness value valid under the iid assumption for each classification set;

a comparative device for selecting the classification set to which the most likely allocated potential classification for the at least one unclassified example belongs, wherein the predicted classification output by the output terminal is the most likely allocated classification according to the strangeness values assigned by the assay means; and

a strength of prediction monitoring device for determining a confidence value for the predicted classification on the basis of the strangeness value assigned by the assay means to one of the classification sets to which the second most likely allocated potential classification of the at least one unclassified example belongs.

2. Data classification apparatus as claimed in claim 1, wherein the processor further includes an example valuation device which determines individual strangeness values for each training classified example and the at least one unclassified example having an allocated potential classification.

3. Data classification apparatus as claimed in claim 2, wherein Lagrange multipliers are used to determine the individual strangeness value.

4. Data classification apparatus as claimed in claim 2, wherein the assay means determines a strangeness value for each classification set in dependence on the individual strangeness values of each example.

5. Data classification apparatus comprising:

an input device for receiving a plurality of training classified examples and at least one unclassified example;

a memory for storing the classified and unclassified examples;

stored programs including an example classification program;

an output terminal for outputting a predicted classification for the at least one unclassified example; and

a processor controlled by the stored programs for identifying the predicted classification of the at least one unclassified example wherein the processor includes:

classification allocation means for allocating potential classifications to each unclassified example and for generating a plurality of classification sets, each classification set containing the plurality of training classified examples and the at least one unclassified example with its allocated potential classification;

assay means for determining a strangeness value valid under the iid assumption for each classification set;

a comparative device for selecting the classification set to which the most likely allocated potential classification for the at least one unclassified example belongs, wherein the predicted classification output by the output terminal is the most likely allocated potential classification according to the strangeness values assigned by the assay means and

a strength of prediction monitoring device for determining a confidence value for the predicted classification on the basis of the strangeness value assigned by the assay means to one of the classification sets to which the second most likely allocated potential classification of the at least one unclassified example belongs.

6. A data classification method comprising: inputting a plurality of training classified examples and at least one unclassified example; identifying a predicted classification of the at least one unclassified example which includes, allocating potential classifications to each unclassified example; generating a plurality of classification sets, each classification set containing the plurality of training classified examples and the at least one unclassified example with its allocated potential classification; determining a strangeness value valid under the iid assumption for each classification set; selecting the classification set to which the most likely allocated potential classification for the at least one unclassified example belongs, wherein the predicted classification is the most likely allocated potential classification in dependence on the strangeness values; determining a confidence value for the predicted classification on the basis of the strangeness value assigned to one of the classification sets to which the second most likely allocated potential classification for the at least one unclassified example belongs; and outputting the predicted classification for the at least one unclassified example and the confidence value for the predicted classification.

7. A data classification method as claimed in claim 6, further including determining individual strangeness values for each training classified example and the at

least one unclassified example having an allocated potential classification.

8. A data classification method as claimed in any one of the preceding claims, wherein the selected classification set is selected without the application of any general rules determined from the training set.

9. A data carrier on which is stored a classification program for classifying data by performing the following steps:

generating a plurality of classification sets, each classification set containing a plurality of training classified examples and at least one unclassified example that has been allocated a potential classification;

determining a strangeness value valid under the iid assumption for each classification set;

selecting the classification set to which the most likely allocated potential classification for the at least one unclassified example belongs, wherein the predicted classification is the most likely allocated potential classification in dependence on the strangeness values; and

determining a confidence value for the predicted classification on the basis of the strangeness value assigned to one of the classification sets to which the second most likely allocated potential classification for the at least one unclassified example belongs.

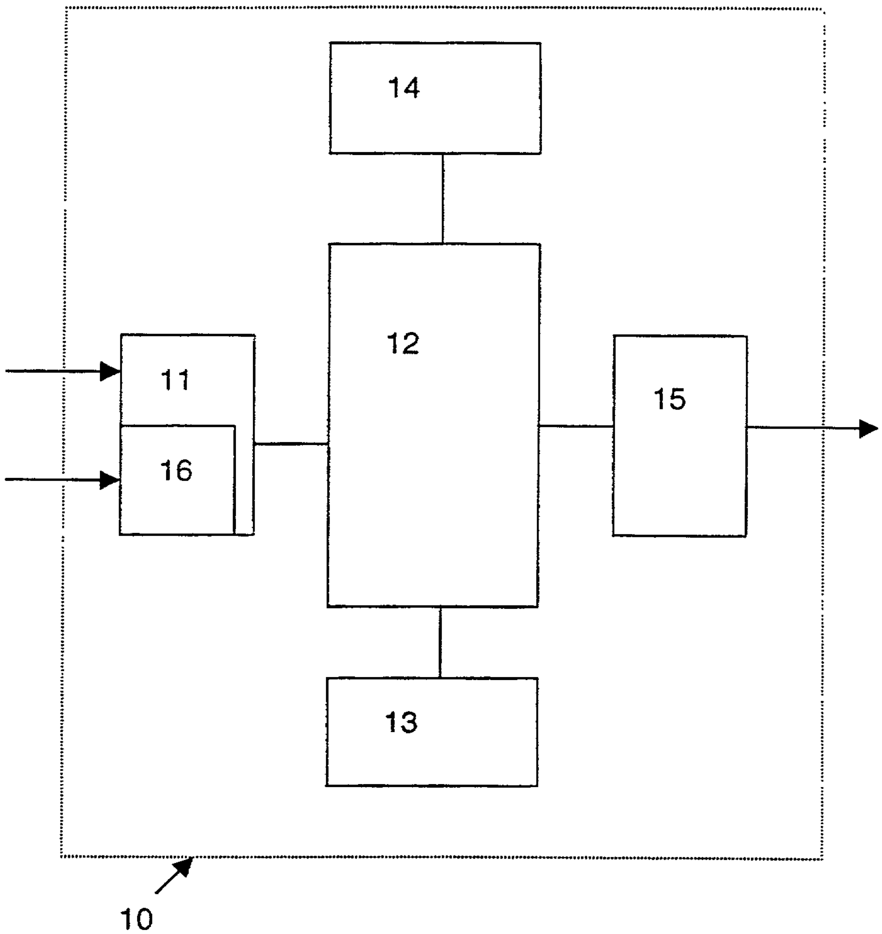

Figure 1

	Training Set					Test Set	
Example No.	1	2	3	4	5	1	2
Example	1	7	1	7	7	7	1
Classification	1	7	1	7	7	?	?

Figure 3

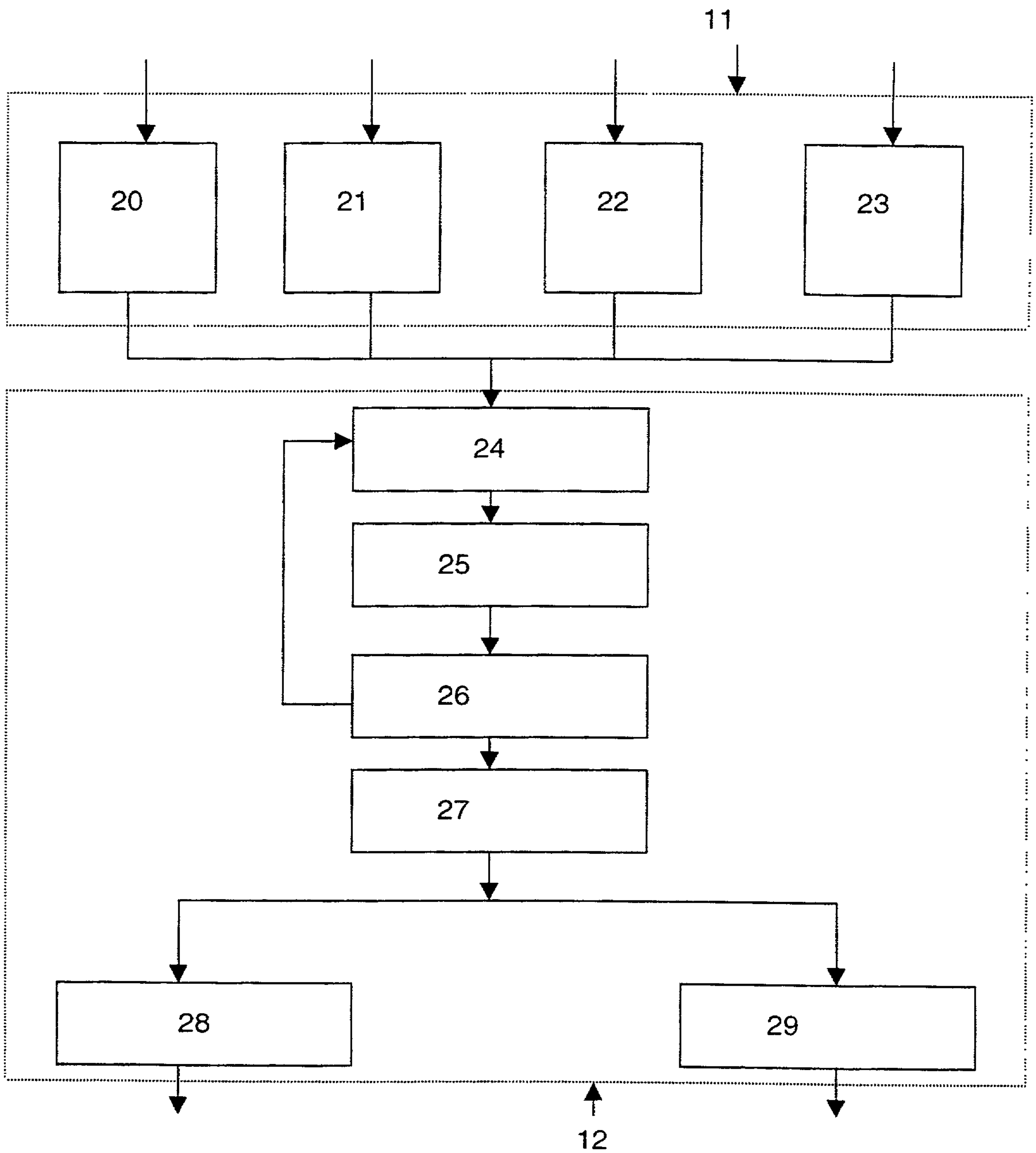

Figure 2

	Test Set									
Example No.	1	2	3	4	5	6	7	8	9	10
Example	1	7	7	1	7	1	1	1	7	7
True Class	1	7	7	1	7	1	1	1	7	7
Predicted Class	1	7	7	1	7	1	1	1	7	7
Confidence	95%	95%	95%	95%	95%	95%	95%	95%	95%	95%
Credibility	19%	100%	100%	100%	100%	28%	100%	100%	100%	100%

Figure 4

COMBINED DECLARATION AND POWER OF ATTORNEY

As below named inventor, I hereby declare that

This declaration is of the following type:

☐ original ☐ design ☐ supplemental
☒ national stage of PCT
☐ divisional ☐ continuation ☐ continuation-in-part

My residence, post office address, and citizenship are as stated below next to my name. I believe I am the original, first, and sole inventor *(if only one name is listed below)* or an original, first, and joint inventor *(if plural names are listed below)* of the subject matter which is claimed and for which a patent is sought on the invention entitled:

DATA CLASSIFICATION APPARATUS AND METHOD THEREOF

the specification of which:

☐ is attached hereto.
☐ was filed on ______________ as Application No. ____________________ and was amended on ________________ *(if applicable)*.
☐ was filed by Express Mail No. ________________, *as Application No. not known yet*, and was amended on ______________ *(if applicable)*.
☒ was described and claimed in PCT International Application No. PCT/GB99/03737 filed on 09.11.99 and as amended pursuant to PCT Article 19 on 31/10.00 *(if any)*.

I state that I have reviewed and understand the contents of the above-identified specification, including the claim(s), as amended by any amendment referred to above.

I acknowledge the duty to disclose information that is material to the patentability of this application in accordance with 37 C.F.R. § 1.56.

I claim foreign priority benefits under 35 U.S.C. § 119 of any foreign application(s) for patent or inventor's certificate or of any PCT international application(s) designating at least one country other than the United States of America listed below and have also identified below any foreign application(s) for patent, utility model, design registration, or inventor's certificate or any PCT international application(s) designating at least one country other than the United States of America filed by me on the same subject matter having a filing date before that of the application(s) of which priority is claimed.

PRIOR FOREIGN PATENT, UTILITY MODEL, AND DESIGN REGISTRATION APPLICATIONS						
COUNTRY	APPLICATION	DATE OF FILING (day,**month**,year)	PRIORITY CLAIMED UNDER 35 U.S.C. § 119			
GB	9824552.5	09.11.98	X	YES		NO
				YES		NO
				YES		NO

I claim the benefit pursuant to 35 U.S.C. § 119(e) of the following United States provisional application(s):

PRIOR U.S. PROVISIONAL APPLICATIONS BENEFIT CLAIMED UNDER 35 U.S.C. 119(e)	
APPLICATION NO.	DATE OF FILING (day,**month**,year)

I claim the benefit pursuant to 35 U.S.C. § 120 of any United States application(s) or PCT international application(s) designating the United States of America listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in that/those prior application(s) in the manner provided by the first paragraph of 35 U.S.C. § 112, I acknowledge the duty to disclose material information as defined in 37 C.F.R. § 1.56 effective between the filing date of the prior application(s) and the national or PCT international filing date of this application.

PRIOR U.S. APPLICATIONS OR PCT INTERNATIONAL PATENT APPLICATIONS DESIGNATING THE U.S. FOR BENEFIT UNDER 35 U.S.C. 120				
U.S. APPLICATIONS		Status *(check one)*		
APPLICATION NO.	U.S. FILING DATE	PATENTED	PENDING	ABANDONED
1. 0 /				
2. 0 /				
3. 0 /				

PCT APPLICATIONS DESIGNATING THE U.S.			Status *(check one)*		
PCT APPLICATION NO.	PCT FILING DATE (day,month,year)	U.S. APPLN. NOS. ASSIGNED (if any)	PATENTED	PENDING	ABANDONED
4.PCT/GB99/03737	09.11.99				
5.					
6.					

DETAILS OF FOREIGN APPLICATIONS FROM WHICH PRIORITY CLAIMED UNDER 35 U.S.C. §119 FOR ABOVE LISTED U.S./PCT APPLICATIONS				
ABOVE APPLN. NO.	COUNTRY	APPLICATION NO.	DATE OF FILING (day,month,year)	DATE OF ISSUE (day,month,year)
1. PCT/GB99/03730	GB	9824552.5	09.11.98	
2.				
3.				
4.				
5.				
6.				

As a named inventor, I hereby appoint Leydig, Voit & Mayer, Ltd. to prosecute this application and transact all business in the Patent and Trademark Office connected therewith: Customer Number 23460.

23460

PATENT TRADEMARK OFFICE

I further direct that correspondence concerning this application be directed to Leydig, Voit & Mayer, Ltd.: Customer Number 23460.

23460

PATENT TRADEMARK OFFICE

I declare that all statements made herein of my own knowledge are true, that all statements made on information and belief are believed to be true, that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code, and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

1-00

Full name of **sole or first inventor:** Alex Gammerman

Inventor's signature A. Gammerman

Date 4/05/2001 Country of Citizenship: UK

Residence: Surrey, United Kingdom GBX
(City, State or Province, and Country)

Post Office Address: c/o Royal Holloway University of London
Department of Computer Science
Egham
Surrey
TW20 0EX
United Kingdom
(Number and Street, City, State or Province, Postal Code, Country)

2-00

Full name of **second joint inventor**, if any: Volodya Vovk

Inventor's signature V Vovk

Date 4/5/2001 Country of Citizenship: Russian Federation

Residence: Surrey, United Kingdom GBX
(City, State or Province, and Country)

Post Office Address: c/o Royal Holloway University of London
Department of Computer Science
Egham
Surrey
TW20 0EX
United Kingdom
(Number and Street, City, State or Province, Postal Code, Country)